LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Transformer Model for Genome Sequence Analysis

BA Thesis Disputation

Supervisors: Prof. Dr. Mina Rezaei, Hüseyin Anil Gündüz, Martin Binder

Noah Hurmer • noah.hurmer@campus.lmu.de • 08.08.2022

Outline

- Motivation
- BERT
- DNABERT Adaptation
- Phage/Non-Phage Task
- Scaled Self-Supervised Trials
- Model Setups
- Results
- Conclusion & Outlook

Motivation

- annotation/analysis of genomes via expensive experimentation
- analyzing read-level length DNA sequences collected from an environment
- NGS realizes abundance of available unlabeled genome sequences

Motivation

- annotation/analysis of genomes via expensive experimentation
- analyzing read-level length DNA sequences collected from an environment
- NGS realizes abundance of available unlabeled genome sequences

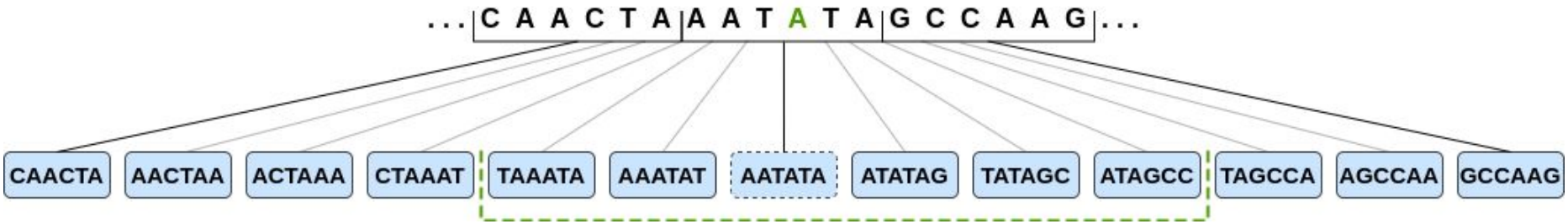
➡ application of semi-supervised approaches to genomic tasks

  - self-supervised training through representation learning
  - adapting methods developed for NLP

DNABERT (2)

- adaptation of *BERT* (1)
  - Transformer Encoder (3) (12 blocks)
  - bidirectional self-attention
  - pretext task of MLM
  - representation size of 768
- preprocessing
  - sentences: sample subsequences from genomes
  - words: tokenize sequences through *k*-mer representation
- mask *k* consecutive tokens during self-supervised pretraining

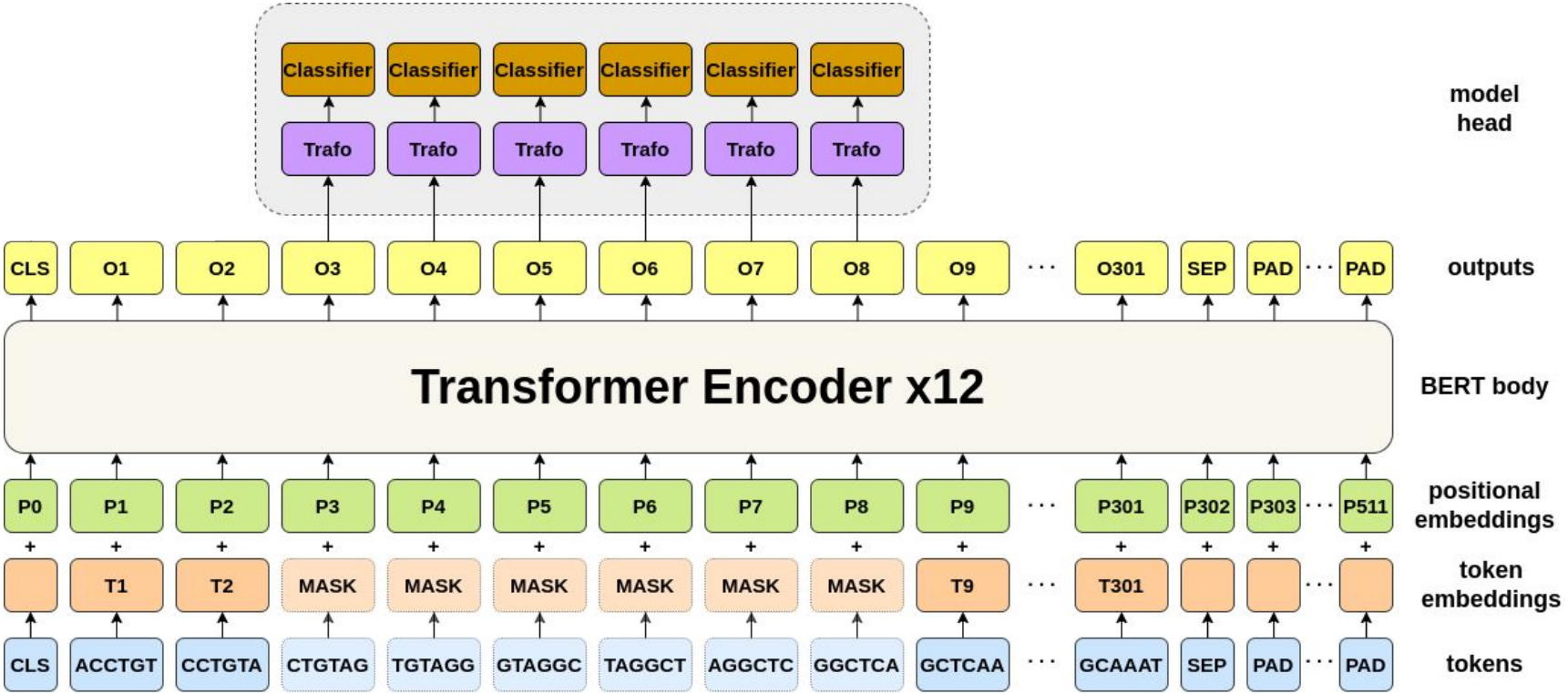
K-mer Creation



DNABERT (2)

- adaptation of *BERT* (1)
  - Transformer Encoder (3) (12 blocks)
  - bidirectional self-attention
  - pretext task of MLM
  - representation size of 768
- preprocessing
  - sentences: sample subsequences from genomes
  - words: tokenize sequences through *k*-mer representation
- mask *k* consecutive tokens during self-supervised pretraining

DNABERT


Classifier
Trafo
model head
CLS O1 O2 O3 O4 O5 O6 O7 O8 O9 ... O301 SEP PAD ... PAD
outputs
Transformer Encoder x12
BERT body
P0 P1 P2 P3 P4 P5 P6 P7 P8 P9 ... P301 P302 P303 ... P511
positional embeddings
T1 T2 MASK T9 T301
token embeddings
CLS ACCTGT CCTGTA CTGTAG TGTAGG GTAGGC TAGGCT AGGCTC GGCTCA GCTCAA GCAAAT SEP PAD PAD
tokens

DNABERT (2)

- adaptation of *BERT* (1)
  - Transformer Encoder (3) (12 blocks)
  - bidirectional self-attention
  - pretext task of MLM
  - representation size of 768
- preprocessing
  - sentences: sample subsequences from genomes
  - words: tokenize sequences through *k*-mer representation
- mask *k* consecutive tokens during self-supervised pretraining

Phage/Non-Phage Task

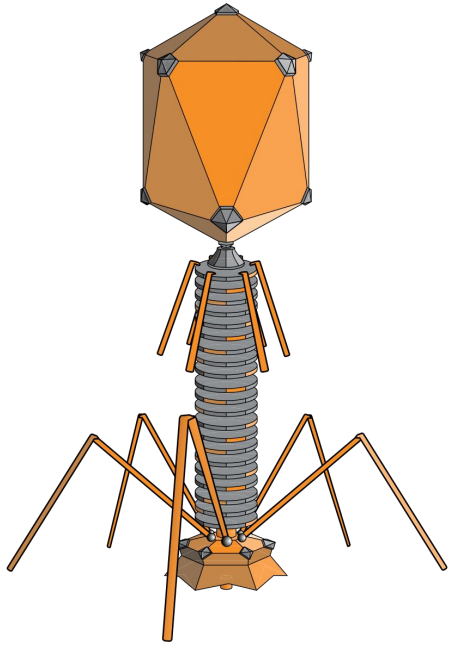
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- collection of virus genomes from *GenBank* (5)
  - ~40k FASTA-files
- self-supervised training on 70% of files
- supervised training with different label availability scenarios (10%, 1%, 0.1%)
- binary classification task of a genome level class
  - identification of bacteriophages
  - read-length sequences 150 and 1000 nucleotides
  - class averaged recall in % and F_1-score
- compare to *self-genomenet* (4) and fully supervised *DNABERT*

What is a Bacteriophage?

- taxonomic classification of viruses
- bacteria as hosts
- uses in medicine (6)
- identification not trivial
  - diverse in genomic organization (7)
  - lack of universal marker genes (8)

figure 1: Adenosine, 2009



Phage/Non-Phage Task

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- collection of virus genomes from *GenBank* (5)
  - ~40k FASTA-files
- self-supervised training on 70% of files
- supervised training with different label availability scenarios (10%, 1%, 0.1%)
- binary classification task of a genome level class
  - identification of bacteriophages
  - read-length sequences 150 and 1000 nucleotides
  - class averaged recall in % and F_1-score
- compare to *self-genomenet* (4) and fully supervised *DNABERT*

Self-Supervised Trials

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- self-supervised training
  - *BERT*-small
  - 10k steps, 10% data
- 36 trials conducted
  - parameters: learning rate, weight decay, warmup %, masking %
  - various masking techniques, k-mer creation and sequence sampling methods
- supervised training
  - phage/non-phage task with 150nt inputs
  - frozen representation layers (linear evaluation)

Self-Supervised Setup I

virBERT

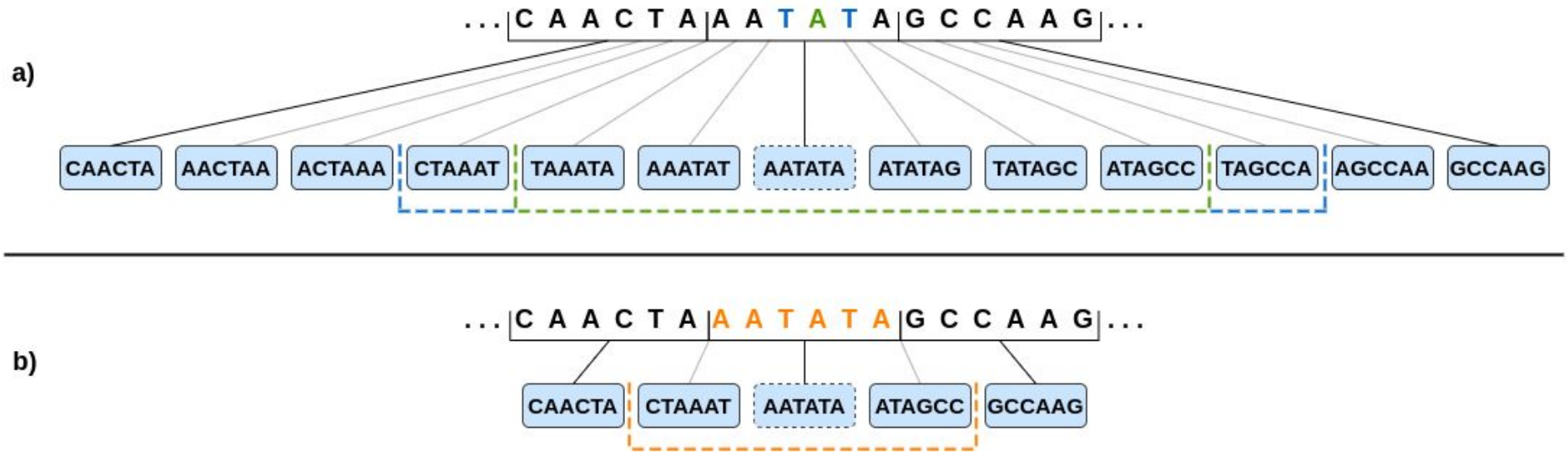
- recreation of *DNABERT* setup
- 5-510nt long sequences tokenized to 6-mers
- mask 6 consecutive tokens at 2.5% sampled locations
- LR of $4e^{-4}$
- AdamW with linear warm-up of 5% of steps

Self-Supervised Setup II

virBERT-mask8

- motivation: mask more nucleotides
- 36-510nt long sequences tokenized to 6-mers
- mask 8 consecutive tokens at 1.875% sampled locations
- LR of $1e^{-3}$
- AdamW with linear warm-up of 10% of steps

K-mer Creation



Self-Supervised Setup III

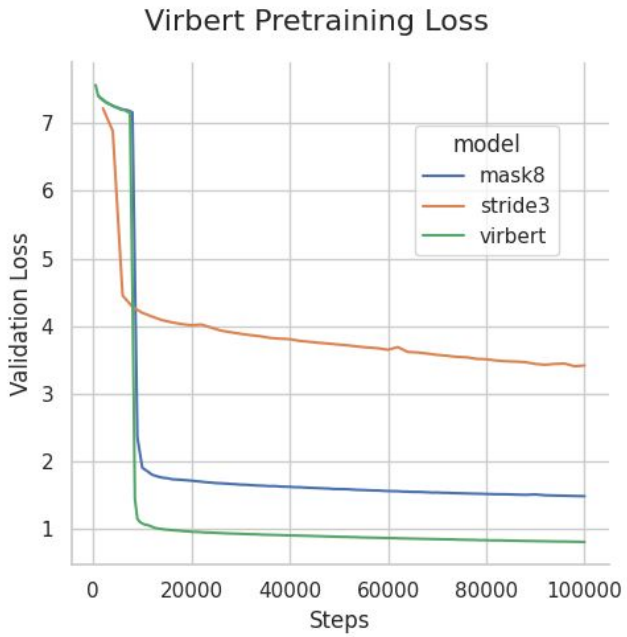
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

virBERT-stride3

- motivation: mask more nucleotides, longer input sequences, faster training
- 36-1000nt long sequences tokenized to 6-mers of stride 3
- mask 3 consecutive tokens at 5% sampled locations
- limited to 340 input tokens
- LR of $1e^{-3}$
- AdamW with linear warm-up of 10% of steps

Self-Supervised Pretraining

- ~8M sequences
- 100k steps
- 8 (5) nvidia A100-40GB GPUs
- 190 (141) hours of training



Supervised Finetuning

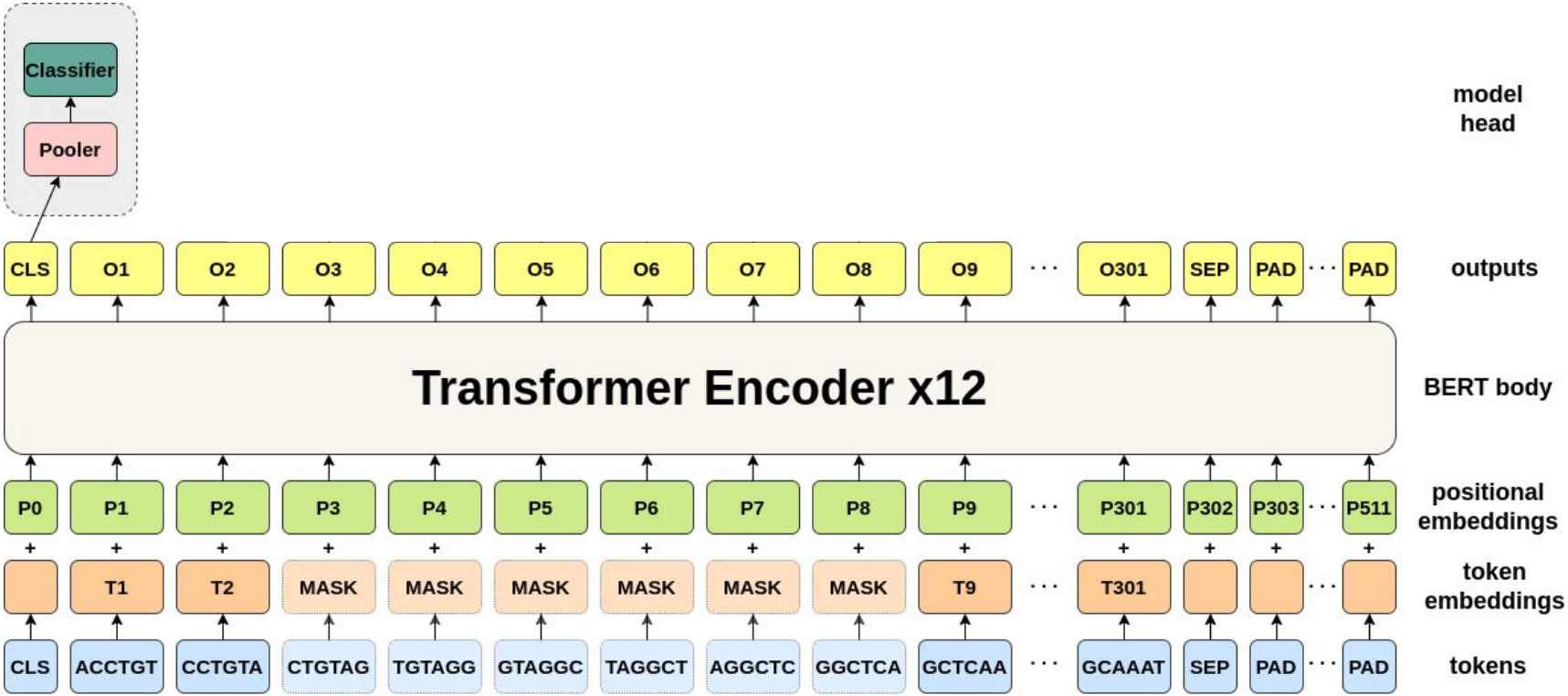
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- sequences sampled randomly from genomes

	10%	1%	0.1%
150nt	8M	2M	500k
1000nt	2M	500k	166k

- training a classifier over the CLS output
- virBERT (-mask8) with input sequences >512
  - split sequences
  - CLS outputs combined with linear layer

DNABERT


Classifier
Pooler
model head
CLS O1 O2 O3 O4 O5 O6 O7 O8 O9 ... O301 SEP PAD ... PAD
outputs
Transformer Encoder x12
BERT body
P0 P1 P2 P3 P4 P5 P6 P7 P8 P9 ... P301 P302 P303 ... P511
positional embeddings
T1 T2 MASK MASK MASK MASK MASK MASK T9 ... T301
token embeddings
CLS ACCTGT CCTGTA CTGTAG TGTAGG GTAGGC TAGGCT AGGCTC GGCTCA GCTCAA ... GCAAAT SEP PAD ... PAD
tokens

Supervised Finetuning

- sequences sampled randomly from genomes

	10%	1%	0.1%
150nt	8M	2M	500k
1000nt	2M	500k	166k

- training a classifier over the CLS output
- virBERT (-mask8) with input sequences >512
  - split sequences
  - CLS outputs combined with linear layer

Bacteriophage Prediction Results

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

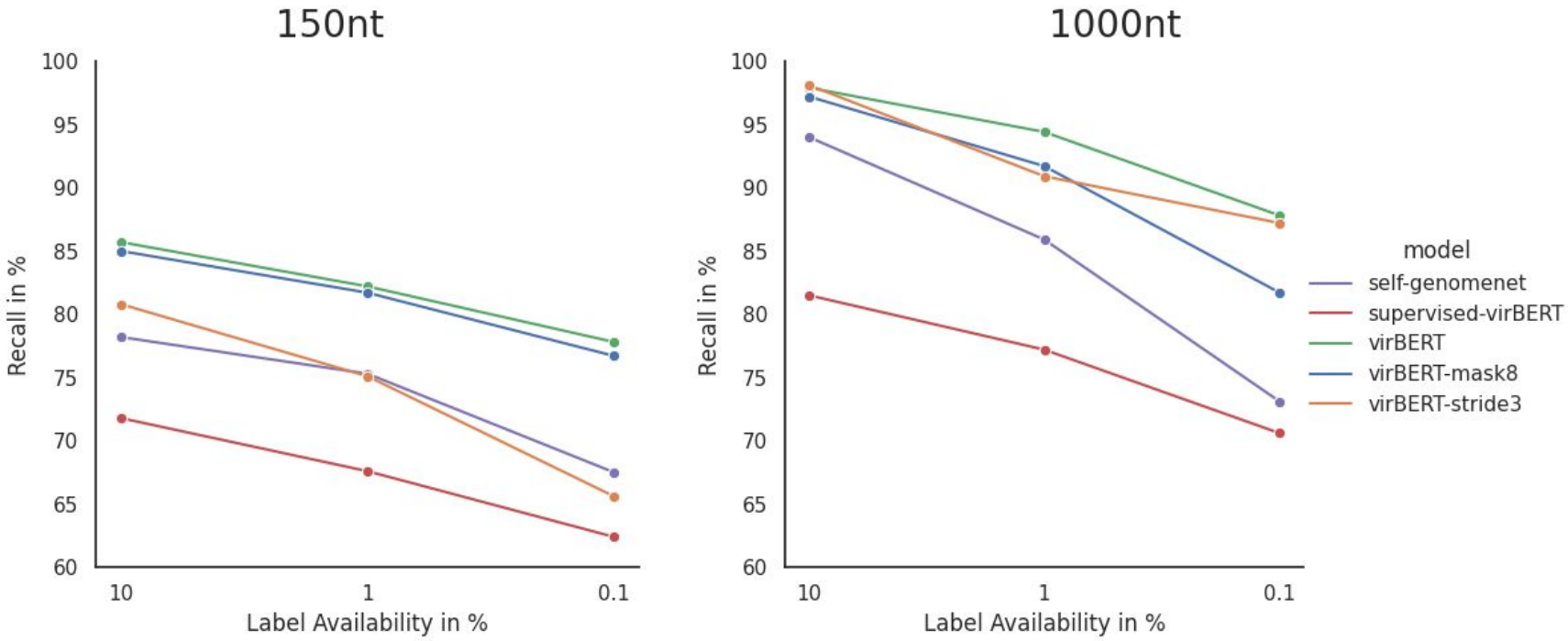
150nt:

	10%		1%		0.1%	
	$Recall_M$	F_1	$Recall_M$	F_1	$Recall_M$	F_1
self-genomenet	78.2	0.785	75.3	0.751	67.2	0.700
supervised-virBERT	71.8	0.710	67.6	0.673	62.4	0.608
virBERT	**85.7**	**0.851**	**82.2**	**0.821**	**77.8**	**0.780**
virBERT-mask8	85.0	0.845	81.7	0.812	76.7	0.762
virBERT-stride3	80.8	0.801	75.1	0.757	65.6	0.654

1000nt:

	10%		1%		0.1%	
	$Recall_M$	F_1	$Recall_M$	F_1	$Recall_M$	F_1
self-genomenet	94.0	-	85.9	-	73.1	0.846
supervised-virBERT	81.5	0.871	77.2	0.867	70.6	0.773
virBERT	97.9	0.986	**94.4**	**0.968**	**87.8**	**0.930**
virBERT-mask8	97.2	0.983	91.7	0.953	81.7	0.901
virBERT-stride3	**98.1**	**0.988**	90.9	0.949	87.2	0.927

Bacteriophage Prediction Results



Conclusion

- *DNABERT* approach outperforms baseline on this task
- gained representations during pretraining contribute considerably
  - 20% $recall_M$ increase over fully supervised model
  - higher accuracy when pretrained on the same type of data regardless of label availability
- base *virBERT* most accurate self-supervised setup
  - masking more nucleotides does not seem advantageous
  - scaled trail results are not directly transferable
- *stride3* variant on par for 1000nt task
  - pretraining and finetuning less resource intensive

Outlook

- improving *stride3* variant
  - masking less tokens
  - data preprocessing
  - more thorough HPO
- validate model's performance
  - additional tasks of different levels & different organisms
  - baselines with the same representation size
- less specific pretraining
- different approaches to creating tokens from genome sequences

References

1. J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. URL https://arxiv.org/abs/1810.04805
2. Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
4. H. A. Gündüz, M. Binder, X.-Y. To, R. Mreches, P. C. Münch, A. C. McHardy, B. Bischl, and M. Rezaei. Self-genomenet: Self-supervised learning with reverse-complement context prediction for nucleotide-level genomics data, 2022. URL https://openreview.net/forum?id=92awwjGxIZI.
5. E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi. GenBank. *Nucleic Acids Research*, 48(D1):D84–D86, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz956. URL https://doi.org/10.1093/nar/gkz956
6. P. Domingo-Calap and J. Delgado-Martínez. Bacteriophages: protagonists of a post-antibiotic era. 2018. *Antibiotics*, 7(3), p.66.
7. P. Simmonds and P. Aiewsakun. Virus classification – where do you draw the line? *Arch Virol*. 2018 Aug, 2018
8. S. Roux, D. P´aez-Espino, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, T. B. K. Reddy, S. Nayfach, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, and N. C. Kyrpides. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, 49(D1):D764–D775, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa946. URL https://doi.org/10.1093/nar/gkaa946

figure 1: (CC BY-SA 3.0) Adenosine. Artistic rendering of a T4 bacteriophage. 2009. Bacteriophage. Retrieved August 4, 2022, from https://en.wikipedia.org/wiki/Bacteriophage#/media/File:PhageExterior.svg

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

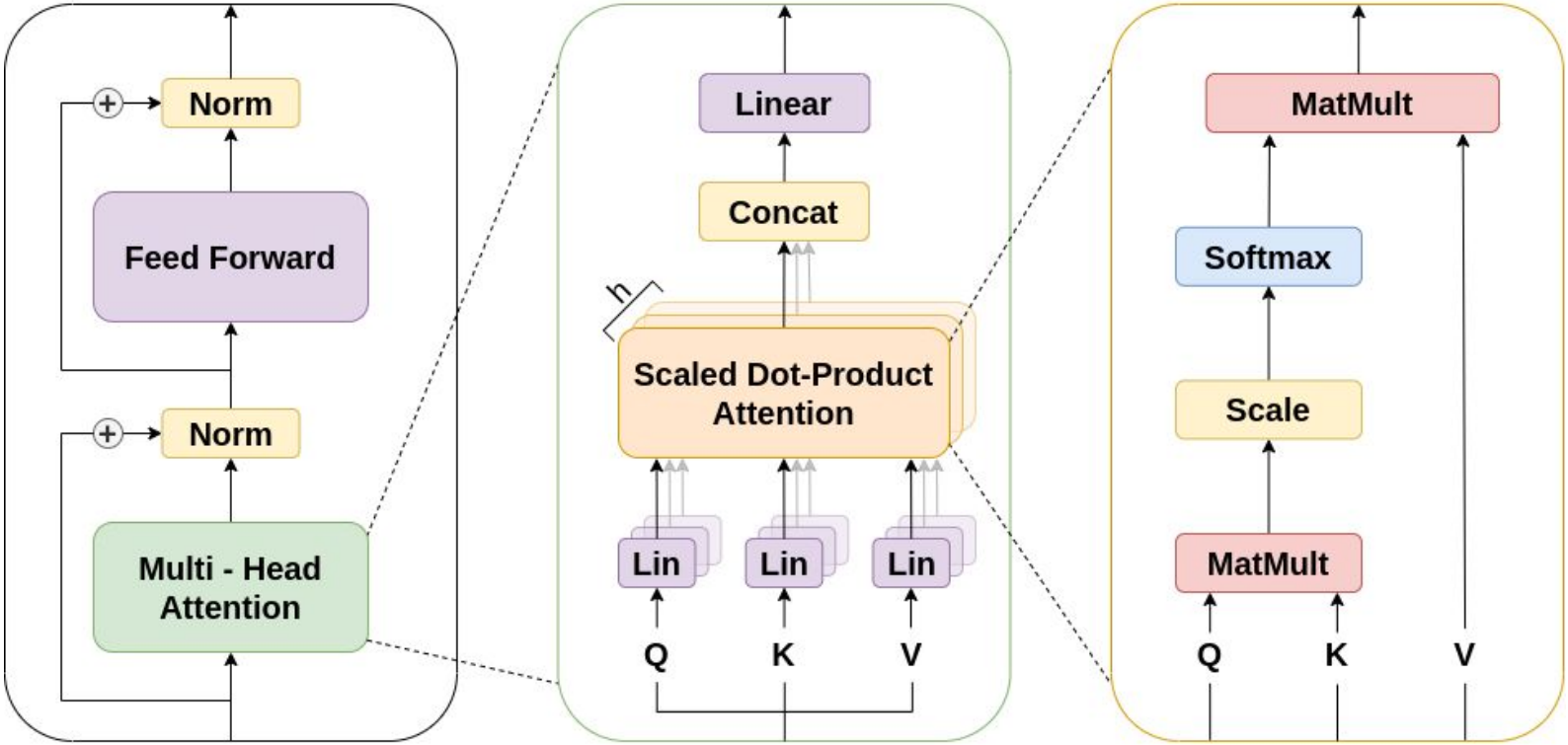
Discussion

Noah Hurmer • noah.hurmer@campus.lmu.de • 08.08.2022

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Appendix

Transformer Encoder Block



Semi-Supervised Learning

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- applications in NLP and CV
- self-supervised pretraining
  - understand meaning, structure and dependencies of the type of data input
  - solve a ‘pretext’ task on unlabeled data
  - learn token representations and model weights
- supervised finetuning on task specific labeled data

Nucleotides hidden during Pretraining

mask %	stride	#nt/loc	%nt hidden	% for k=6
15	1	1	$\frac{15}{k}$	2.5
15	1	3	$\frac{45}{k+2}$	5.625
15	1	5	$\frac{75}{k+4}$	7.5
15	1	6	$\frac{90}{k+5}$	8.1818
15	$\frac{k}{2}$	$\frac{k}{2}$	15	15
15	k	k	15	15

Metrics

- class averaged recall in %

$$Recall_M = 100 \times \frac{\sum_{i=1}^{n} Recall(c_i)}{n}$$

$c_i \in C$ Classes $\qquad$ n number of classes

- F_1-score

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

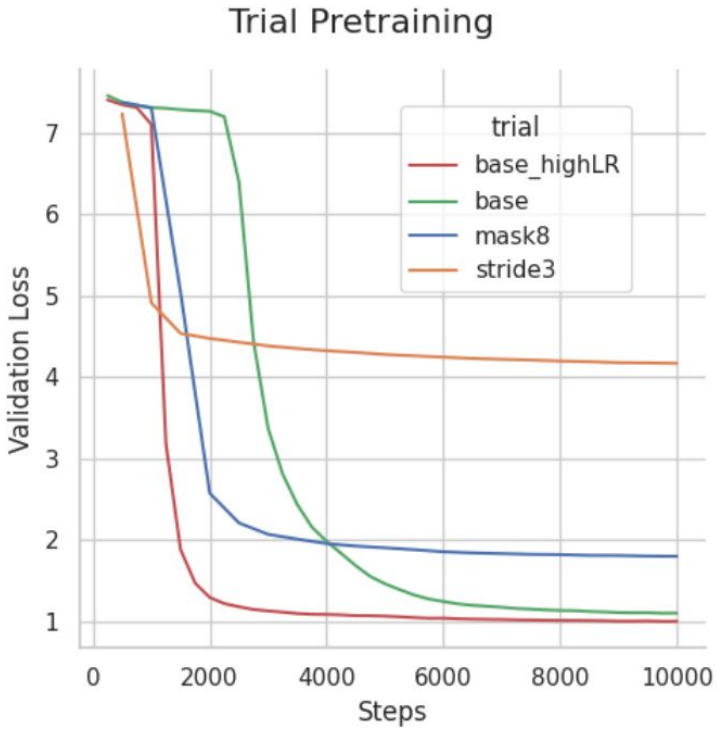
Training Time

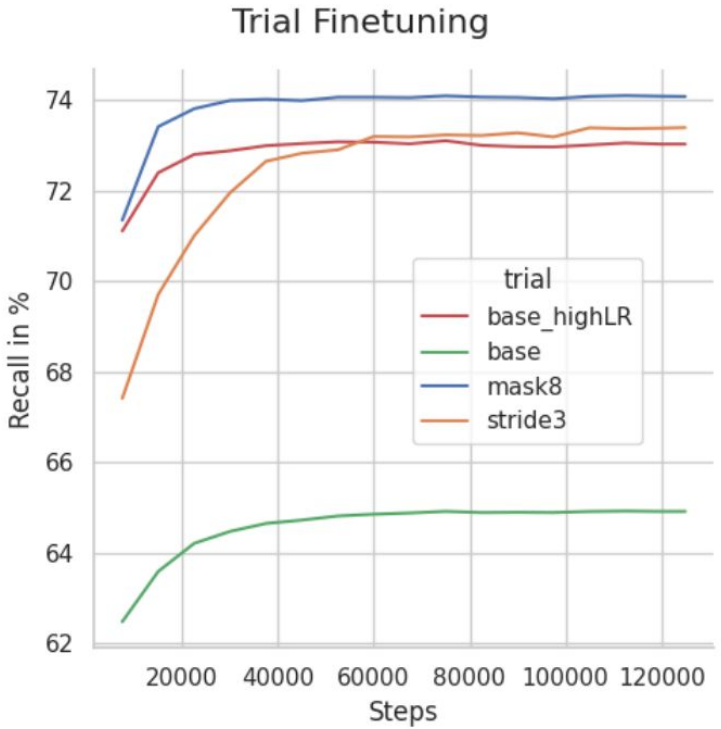
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

model	Training time	#GPUs	GPU Type	accumulated time
self-genomenet	158	-	RTX 2080 Ti	-
virBERT	190	8	A100	1520
virBERT-mask8	190	8	A100	1520
virBERT-stride3	141	5	A100	705

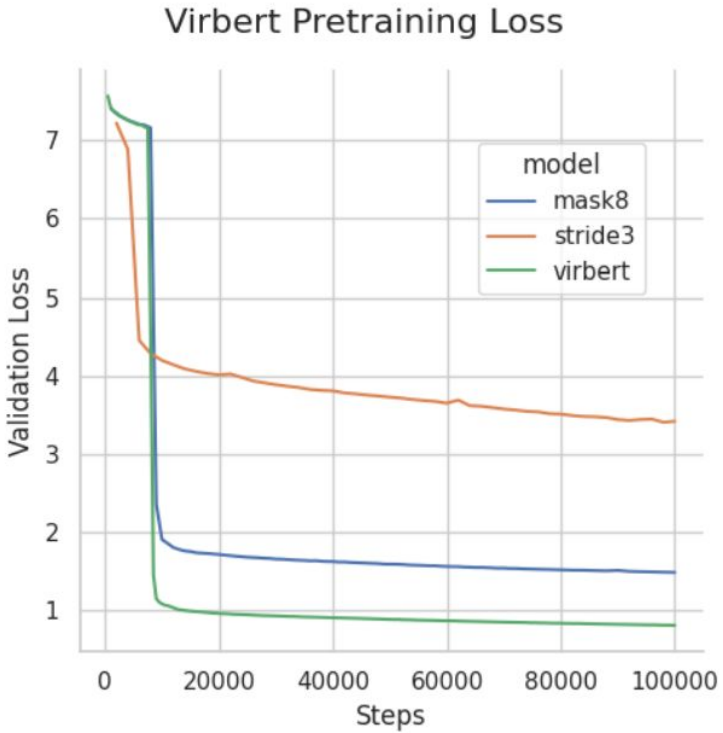
	150nt		1000nt		
model	#GPUs	steps/h	#GPUs	steps/h	steps/h/gpu
virBERT(-mask8)	1	4870	2	3564	1782
virBERT-stride3	1	13158	1	4478	4478

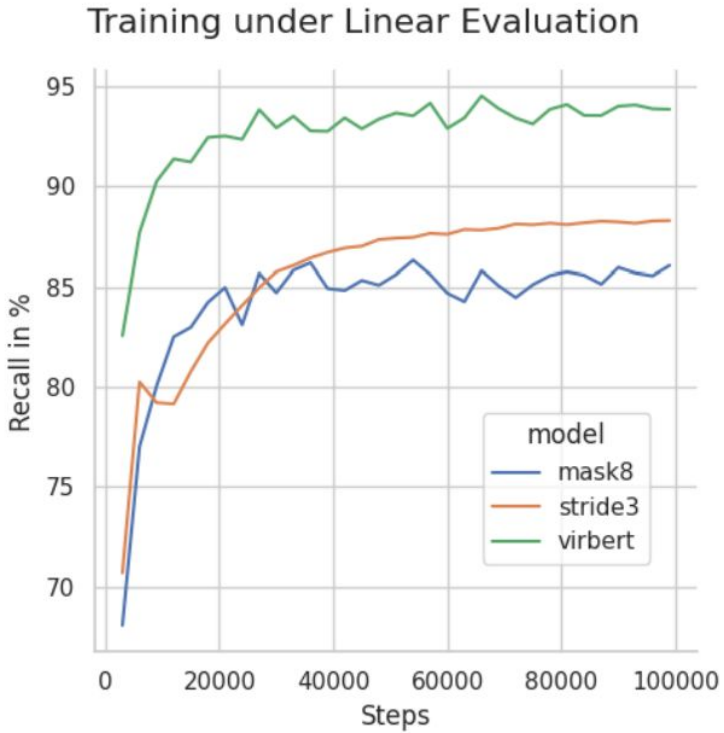
Trial Training





Full Scale Training



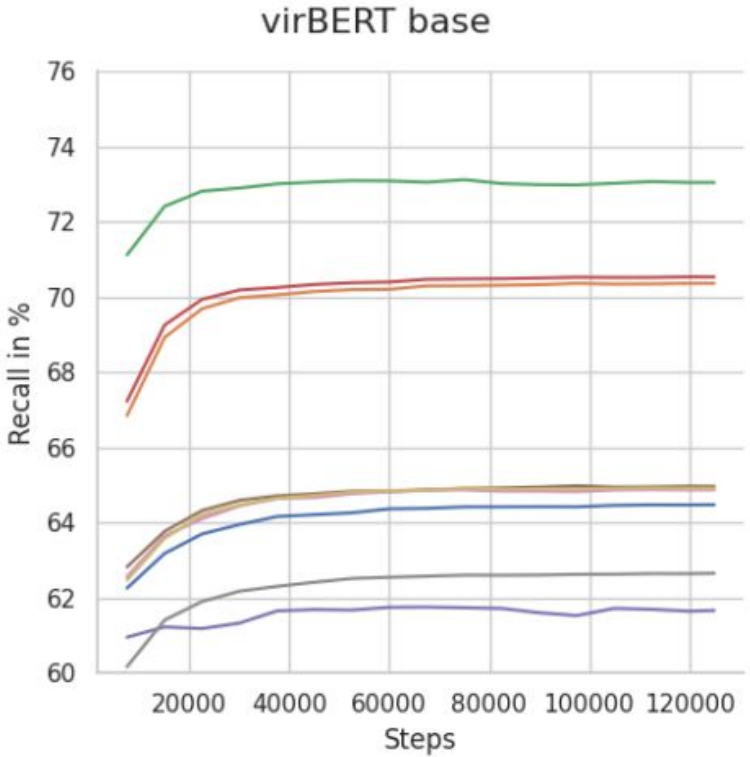


Trial Results

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

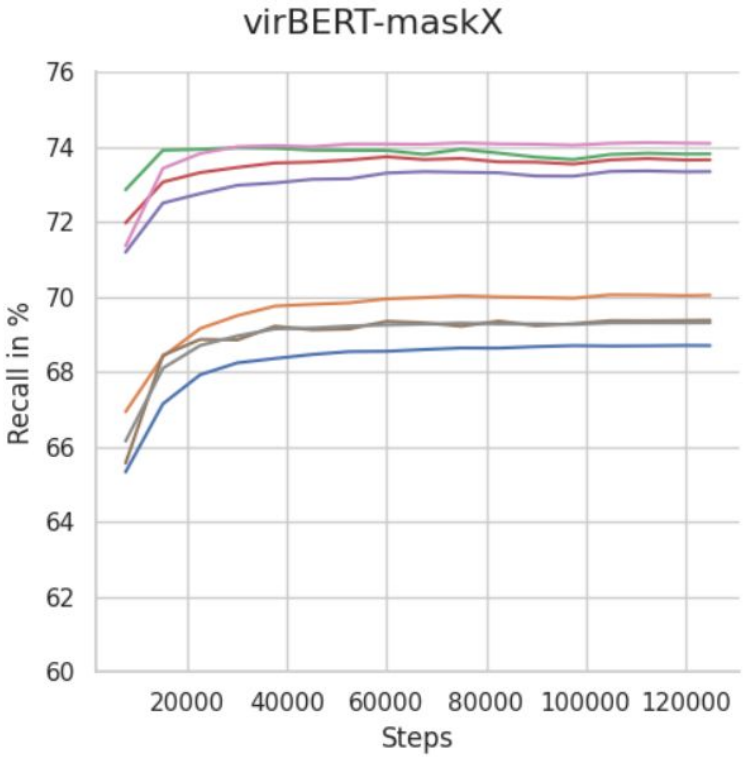
- hyperparameter improvements
  - higher learning rate up to $1e^{-3}$ (compared to $4e^{-4}$)
  - higher warmup-percentage of 10% (5%)
- masking technique
  - masking more than 6 consecutive tokens seems better (not consistent)
  - *mask8* overall the best setup
- k-mers of higher stride
  - generally less accurate, stride of 3 better than 6
  - more similar pretraining sequence lengths to the downstream task may be beneficial

Trials I



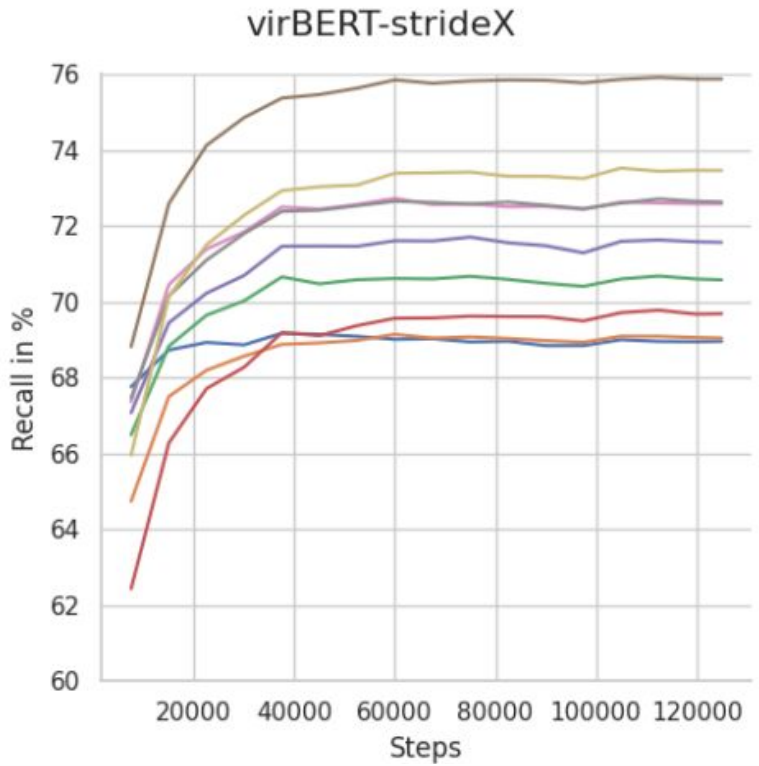
id	description
base	base virBERT
1	mask% 20
2	med lr
3	low lr
5	lower weight decay
8	higher low_b
9	higher ratio
12	med lr, fluid mask%
13	high lr

Trials II



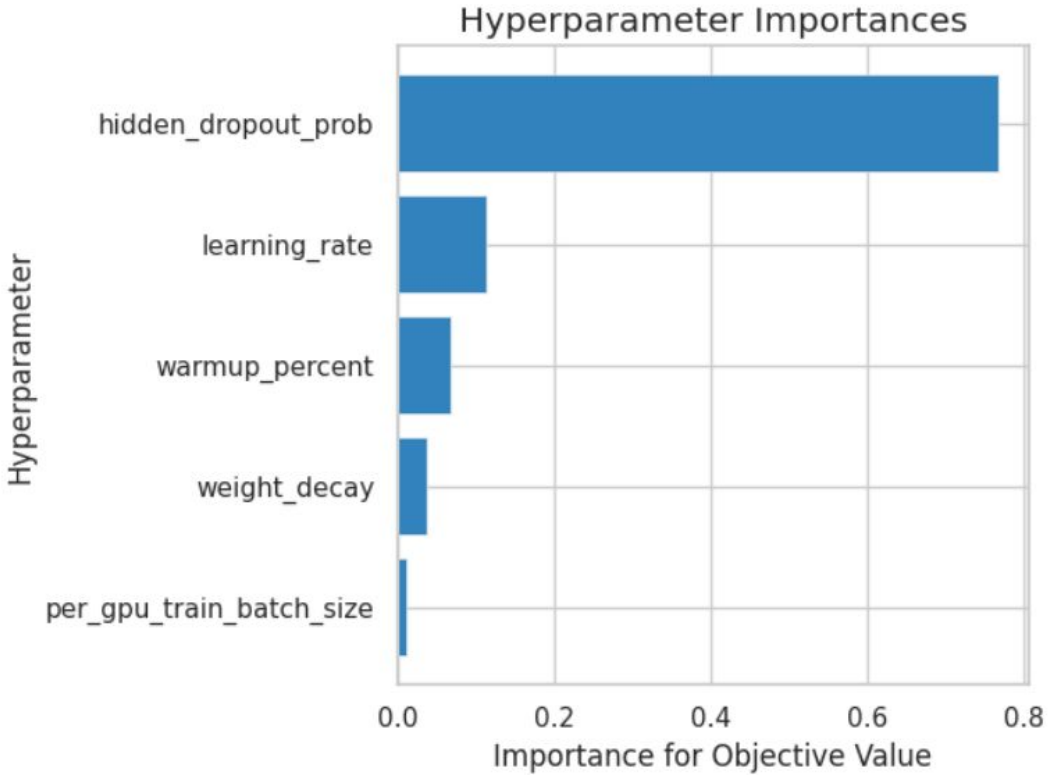
id	description
7	m8, med lr
14	m8, med lr, mask% 20
15	m10, med lr
19	m8, high lr
20	m10, high lr
21	m11, high lr
26	m8, lr $2e^{-3}$
30	= 19 but warm-up% 10

Trials III



id	description
11	s3, base lr
16	s6, base lr
17	s3, high lr
18	s6, high lr
25	s3, upp_b 1000nt
28	s3, upp_b 510nt, cap
29	s3, upp_b 510nt
31	s3, upp_b 1000nt, cap
33	= 31 but lower bias

HPO


Hyperparameter Importances
hidden_dropout_prob
learning_rate
warmup_percent
weight_decay
per_gpu_train_batch_size
Hyperparameter
0.0
0.2
0.4
0.6
0.8
Importance for Objective Value

Linear Evaluation

- 1000nt input sequences

	$Recall_M$	F_1
self-genomenet	88.6	0.916
virBERT	97.8	0.986
virBERT-mask8	93.8	0.963
virBERT-stride3	92.6	0.956

Transfer Learning

- Phage/Non-Phage task
- models pretrained with sequences of different biological classification
- 100% label availability

		150nt		1000nt	
model	pt-data	$Recall_M$	F_1	$Recall_M$	F_1
DNABERT6	human	79.2	0.799	96.6	0.978
self-genomenet	bacteria	-	-	97.0	-